Clinical bleeding and thrombin generation in admissions to critical care with prolonged prothrombin time: an exploratory study

Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease.

J. Kenneth Baillie[1,2,3]*, Andrew Bretherick[1] Christopher S. Haley[1,4], Sara Clohisey[1], Alan Gray[5], Lucile P. A. Neyton[1], Jeffrey Barrett[6], Eli A. Stahl[7], Albert Tenesa[1], Robin Andersson[8], J. Ben Brown[9], Geoffrey J. Faulkner[10], Marina Lizio[11], Ulf Schaefer[12], Carsten Daub[11], Masayoshi Itoh[11,13], Naoto Kondo[11], Timo Lassmann[11], Jun Kawai[11], IIBDGC Consortium, FANTOM5 Consortium Damian Mole[2], Vladimir B. Bajic[14], Peter Heutink[15], Michael Rehli[16], Hideya Kawaji[11,13], Albin Sandelin[8], Harukazu Suzuki[11], Jack Satsangi[4], Chrstine A. Wells[16], Nir Hacohen[17], Thomas C Freeman[1], Yoshihide Hayashizaki[11], Piero Carninci[11], Alistair R.R. Forrest[18]*, David A. Hume[1,10]*

1. Division of Genetics and Genomics, The Roslin Institute, University of Edinburgh, Edinburgh, UK

2. Centre for Inflammation Research, University of Edinburgh, Edinburgh, UK

3. Intensive Care Unit, Royal Infirmary Edinburgh, Edinburgh, UK

4. Institute for Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

5. Edinburgh Parallel Computing Centre, The University of Edinburgh, Edinburgh, UK

6. Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

7. Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, USA

8. The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark

9. Department of Statistics, University of California, Berkeley, USA

10. Mater Research Institute, University of Queensland, University of Queensland, Brisbane, Australia.

11. RIKEN Omics Science Center, Yokohama, Japan,, Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan

12. Department for Infectious Disease Informatics, Public Health England, Colindale, UK.

13. RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Japan

14. King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center, Kingdom of Saudi Arabia.

15. German Center for Neurodegenerative Diseases, Tübingen, Germany

16. Dept. Hematology, University Hospital Regensburg, 93053 Regensburg

16. Australian Institute for Bioengineering and Nanotechnology, University of Queensland, St Lucia, Brisbane Australia 4072

17. Broad Institute of Harvard and MIT, Cambridge, USA

18. Harry Perkins Institute of Medical Research, and the Centre for Medical Research, University of Western Australia, QEII Medical Centre, Nedlands, Perth, Western Australia 6009, Australia

*to whom correspondence should be addressed (j.k.baillie@ed.ac.uk, alistair.forrest@perkins.uwa.edu.au, david.hume@uq.edu.au)

Abstract

Genetic variants underlying complex traits, including disease susceptibility, are enriched within the transcriptional regulatory elements, promoters and enhancers. There is emerging evidence that regulatory elements associated with particular traits or diseases share similar patterns of transcriptional activity. Accordingly, shared transcriptional activity (coexpression) may help prioritise loci associated with a given trait, and help to identify underlying biological processes. Using cap analysis of gene expression (CAGE) profiles of promoter- and enhancer-derived RNAs across 1824 human samples, we have analysed coexpression of RNAs originating from trait-associated regulatory regions using a novel quantitative method (network density analysis; NDA). For most traits studied, phenotype-associated variants in regulatory regions were linked to tightly-coexpressed networks that are likely to share important functional characteristics. Coexpression provides a new signal, independent of phenotype association, to enable fine mapping of causative variants. The NDA coexpression approach identifies new genetic variants associated with specific traits, including an association between the regulation of the OCT1 cation transporter and genetic variants underlying circulating cholesterol levels. NDA strongly implicates particular cell types and tissues in disease pathogenesis. For example, distinct groupings of disease-associated regulatory regions implicate two distinct biological processes in the pathogenesis of ulcerative colitis; a further two separate processes are implicated in Crohn's disease. Thus, our functional analysis of genetic predisposition to disease defines new distinct disease endotypes. We predict that patients with a preponderance of susceptibility variants in each group are likely to respond differently to pharmacological therapy. Together, these findings enable a deeper biological understanding of the causal basis of complex traits.

Author Summary

We discover that genetic variants associated with specific diseases have more in common with each other than we have previously seen. Specifically, variants associated with the same disease tend to be in parts of the genome that are turned on or off in similar complex patterns across many different cell types. We discover that genetic variants associated with specific diseases are found within regulatory elements that share patterns of expression. Specifically, variants associated with the same disease tend to be in parts of the genome that are turned on or off together in similar complex patterns across many different cell types. Knowing this helps us to find new variants associated with some diseases, and to better understand the genetic causes of other diseases. Furthermore, we discover that the genetic causes of inflammatory bowel disease fall into two distinct patterns, indicating that two aetiologically-distinct endotypes of this condition exist. Unlike other methods to learn about disease mechanisms from genetic information, our approach does not require any knowledge or assumptions about the genes themselves – it depends only on the patterns in which parts of the genome are activated in different cell types.

Introduction

Genome-wide association studies (GWAS) have considerable untapped potential to reveal new mechanisms of disease[1]. Variants associated with disease are over-represented in regulatory, rather than protein-coding, sequence; this enrichment is particularly strong in promoters and enhancers[2–4]. There is emerging evidence that gene products associated with a specific disease participate in the same pathway or process[5], and therefore share transcriptional control[6].

We have recently shown that cell-type specific patterns of activity at multiple alternative promoters[7] and enhancers[3] can be identified using cap-analysis of gene expression (CAGE) to

detect capped RNA transcripts, including mRNAs, lncRNAs and eRNAs[3,5]. In the FANTOM5 project, we used CAGE to locate transcription start sites at single-base resolution and quantified the activity of 267,225 regulatory regions in 1824 human samples (primary cells, tissues, and cells following various perturbations)[8].

Unlike analysis of chromatin modifications or accessibility, the CAGE sequencing used in FANTOM5 combines extremely high resolution in three relevant dimensions: maximal spatial resolution on the genome, quantification of activity (transcript expression) over a wide dynamic range, and high biological resolution – quantifying activity in a much wider range of cell types and conditions than any previous study of regulatory variation[2,4]. Since a majority of human protein-coding genes have multiple promoters[5] with distinct transcriptional regulation, CAGE also provides a more detailed survey of transcriptional regulation than microarray or RNAseq resources. Heritability of traits studied by some GWAS is substantially enriched in these FANTOM5 promoters[9][10].

Genes that are coexpressed are more likely to share common biology[11,12]. Similarly, regulatory regions that share activity patterns are more likely to contribute to the same biological pathways[5]. We have previously shown transcriptional activity of regulatory elements (both promoters and enhancers[3]) is associated with variable levels of expression arising at these elements in different cell types and tissues[5]. Informative regulatory networks can be derived from predicted transcription factor interactions with FANTOM5 regulatory regions[6]. We therefore use transcript expression here as a surrogate for transcriptional regulatory activity.

In contrast to previous studies[6,13,14], we sought to explore the similarities in activity at disease-associated sets of regulatory regions, rather than genes, and independent of transcription factor binding predictions.

In order to determine whether coexpression of regulatory elements can provide additional information to prioritise genome-wide associations that would otherwise fall below genome-wide significance, we developed network density analysis (NDA). The NDA method combines genetic signals (disease association in a GWAS) with functional signals (correlation in promoter and enhancer-associated transcript levels measured by CAGE across numerous cell types and tissues, Fig 1), by mapping genetic signals onto a pairwise coexpression network of regulatory regions, and then quantifying the density of genetic signals within the network. Every expressed regulatory region that contains a GWAS SNP associated with a given trait is assigned a score quantifying its proximity in the network to every other regulatory region containing a GWAS SNP for that trait. We then identified specific cell types and tissues in which there is preferential activity of regulatory elements associated with selected disease-related phenotypes, thereby providing appropriate cell culture models for critical disease processes.

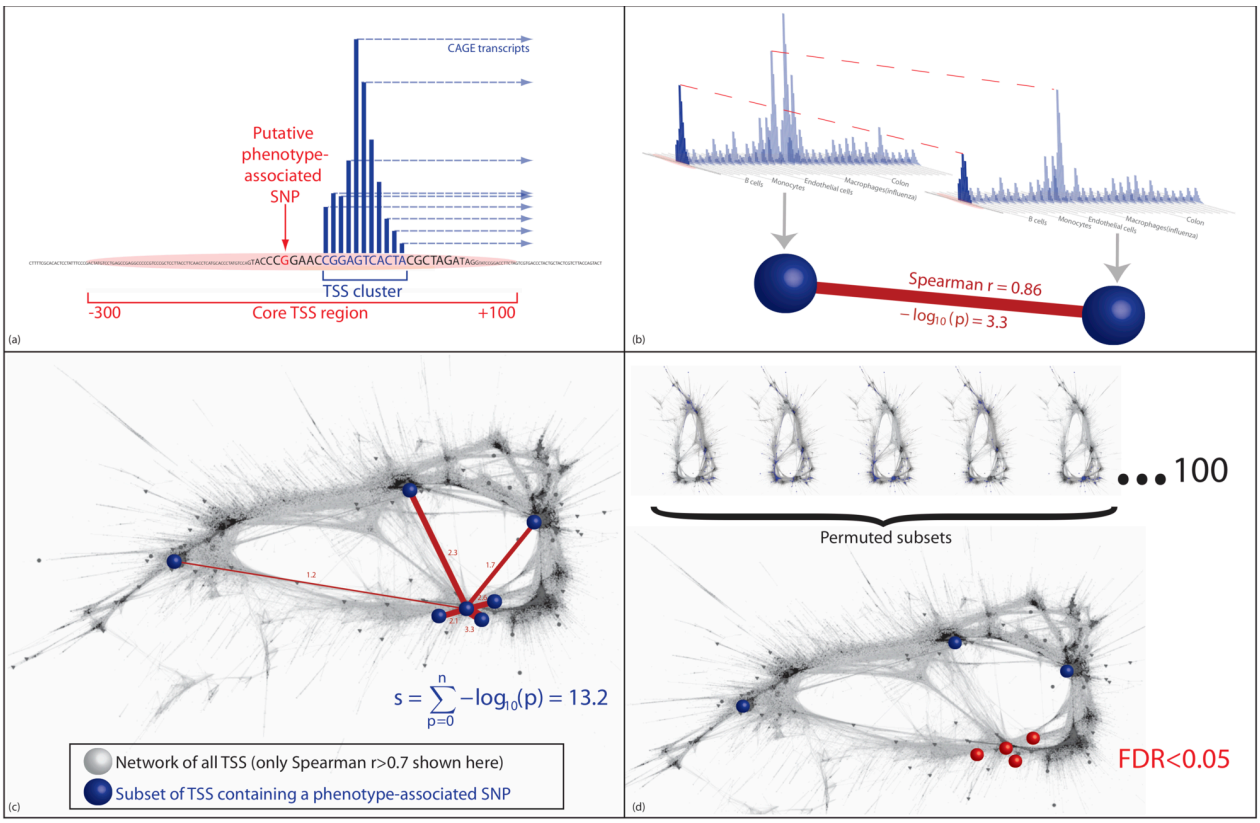

Fig 1. Use of NDA to detect coexpression. (a) A subset of regulatory elements is identified containing disease-associated SNPs. (b) The strength of the links between pairs of these regulatory regions is quantified, first as the Spearman correlation, then as the $-log_{10}$p-value quantifying the probability, specific to this regulatory region, of a Spearman correlation of at least this strength arising by chance. This is determined from the empirical distribution of correlations between this regulatory region and all other regulatory regions in the entire network of all regulatory regions in the genome. (c) The subset of regulatory regions containing disease-associated SNPs form an unexpectedly dense grouping in the network, but this may not be visible in a two-dimensional representation (for illustration, this network shows all correlations between regulatory regions with Spearman $r > 0.7$, layout generated by the FMMM algorithm). The NDA score assigned to any one node is the sum of the links it shares with other nodes in the chosen subset (see Supplementary Methods for a full explanation). d) NDA scores from the input subset of regulatory elements are compared with NDA scores from permuted subsets of regulatory elements in order to quantify the false discovery rate (FDR).

Methods

Regulatory regions

For the purpose of this analysis, promoters identified in the FANTOM5 dataset were defined as the region from -300 bases to +100 bases from a transcription start site[15]. Previous analysis demonstrated that this covers the areas of maximal sequence conservation across species and the core region of transcription factor binding. Enhancers are widely transcribed across the human genome (eRNAs). Since eRNA TSS are considerably longer than promoter TSS (median length(IQR) 272(173-367) vs 15(9-26)), enhancers were defined by the range covered by eRNA transcription start sites.

Coexpression algorithm

For each GWAS study, SNPs were identified that lie within either a functional promoter or enhancer. Any promoter or enhancer that contained a variant putatively associated with a given phenotype was considered to be candidate phenotype-associated regulatory region. A pairwise

matrix was then generated from the full FANTOM5 dataset of promoters and enhancers, in which each node is a regulatory region, and edges reflect the similarity in activity (expression) patterns arising at these regulatory regions, across different cell types and tissues.

To test the hypothesis that regulatory regions genetically associated with a given phenotype are more likely to share activity patterns, we devised the NDA method, which quantifies the strength of coexpression among a chosen pool of putative phenotype-associated regulatory regions. This approach avoids arbitrary cut-offs between clusters (or "communities") of nodes, and yields a single value for each node, quantifying the closeness with all other nodes in a particular subset (network density). NDA was used to integrate the putative association between a regulatory sequence and the phenotype of interest (indicated by the presence of a phenotype-associated SNP), with the coexpression similarity between this node with other nodes that are also putatively associated with the same phenotype.

Principle of network density analysis (NDA)

NDA integrates information from two distinct and independent sources: the relationships between nodes in the network, and the choice of subset. In the present work, nodes are regulatory regions, the subset is those regulatory regions that contain variants associated with a particular phenotype. Spearman's rank correlation was chosen to quantify pairwise relationships, in view of the robustness of this measure in a variety of different distributions. However, the NDA approach is generalisable to any network of pairwise relationships.

Within a network of all possible pairwise relationships between nodes, a subset of nodes is selected that share a particular characteristic. Within this subset of nodes, every pair of nodes is considered. Each relationship between two nodes is expressed as the $-log_{10}$ of the empirical probability of a relationship at least as strong occurring between the chosen node and another, randomly-chosen, node from anywhere in the network. These probabilities are specific to each node and are directional. The NDA score is the sum of the $-log_{10}(p)$ values for a node in the chosen subset and all other nodes within the subset. The NDA score therefore quantifies the density of this subset of nodes in network space. The purpose of using the empirical probability of a correlation, rather than the raw correlation metric, is to control for bias in favour of highly-connected nodes, as would occur if one expression profile were very common. Finally, the NDA score is assigned its own p-value by comparison to that obtained using randomly permuted subsets (see below). If the network contains no additional information about this subset of nodes, then the relationships between nodes in the chosen subset will be no stronger than the relationships seen in permuted subsets.

Application to coexpression of regulatory regions

From the set of all nodes in a network, a subset is selected because they share some characteristic. In the case of the genomic analyses reported here, the nodes are TSS, and the subset of interest is those TSS that contain a variant that has some evidence of association with a particular trait. Throughout this paper, we have defined the set of phenotype-associated transcription start sites, R, as follows: the set of regulatory elements associated with phenotype-associated single nucleotide polymorphism within 300bp (promoters) or 0bp (enhancers) upstream from a FANTOM5 transcription start site (TSS) and 100bp (promoters) or 0bp (enhancers) downstream. In order to enable the detection of new associations, we use a deliberately permissive threshold. We define as "putatively-significant" a SNP-phenotype association of $p < 5\times10^{-6}$. Let the integer variable i be used to index the base pairs (bp) of the genome. For a given trait, the set of input SNPs, K, are those that have a putatively-significant association with that trait

at our chosen threshold. If we let TSS_{start} equal the base pair index 300bp (promoters) or 0bp (enhancers) upstream from a FANTOM5 transcription start site (TSS) and TSS_{end} 100bp (promoters) or 0bp (enhancers) downstream, the set, P, of putative trait-associated promoters is given by:

$$P = \{i : i \in K, TSS_{start} - 300 \leq i \leq TSS_{end} + 100\}$$

and the set E of enhancers containing a putative trait-associated SNP is given by:

$$E = \{i : i \in K, TSS_{start} \leq i \leq TSS_{end}\}$$

giving a total set of regulatory regions:

$$R = P \cup E$$

Linkage disequilibrium (LD) - grouping nearby regulatory regions

Input SNPs from GWAS results tend to be in LD with nearby variants. There is therefore a risk of spurious coexpression, since nearby regulatory regions are also likely to share regulatory influences, such as chromatin accessibility, enhancers, and lncRNAs. One solution to this would be to filter input SNPs by LD. However this would require that LD relationships for all SNPs be known for all of the populations from which SNP association data were derived, which is not the case. It would also risk removing functionally important regulatory regions from the analysis, by choosing only one SNP per LD block.

In order to overcome these problems, we sought to identify those regulatory region-associated SNPs within a given region that are most likely to contribute to a given subnetwork of putative phenotype-associated regulatory regions. By the definitions described above, these will be those regulatory regions with the highest NDA score. Regulatory regions are considered for combination if they are separated by 100,000bp or less. If any regulatory region within this range has a correlation p-value of less than 0.1 with any other regulatory regions in the range, they are combined. A single representative regulatory region is then chosen - the regulatory region with the largest NDA score in the group, derived from a network comprised of all other groups.

In order to confirm that spurious coexpression signals are not being generated solely because of LD, we used the ENSEMBL Perl API for the 1000 genomes phase 3 data (CEU) to search for variants in LD with each SNP lying within the chosen regulatory region for each group. Variants in LD with a variant in any other chosen regulatory region are reported.

Coexpression matrix

A is defined as the set of all nodes in the whole network. Each member of A is a node in an interaction network. For each $i \in R$, Spearman's rank correlation, x, is calculated with each other node in R. The probability, p, of a correlation as strong as, or stronger than, the index correlation, x, arising by a chance pairing between the index node and any other node ($n_{(r>x)}$) is inferred from the empirical distribution of all correlations (r) of the index node in A.

$$p = \frac{n_{(r>x)}}{n_A}$$

Network density analysis

For every node in the set R, a score s is calculated to summarise the strength of interactions with all other nodes in R. Since the only thing that the elements of R have in common is that they are

TSS identified by the set of input SNPs, unexpectedly strong inter-relationships between elements of R are taken as indirect evidence of a relationship between the input SNPs themselves. The NDA score, s, is defined as the sum of $-log_{10}(p)$ values for interaction strength within the matrix.

$$s = \sum_{p=0}^{n} -log_{10}(p)$$

Raw p-values are calculated from the empirical distribution of values of s for 10000 permuted networks. The Benjamini-Hochberg method is used to estimate false discovery rate (FDR). Significant network density scores are taken as those with $FDR < 0.05$. In order to enable comparison of coexpression scores between different analyses, the raw coexpression score (s) is corrected by dividing by the total number of independent groups of regulatory regions included in each analysis, n_{res}, yeilding a corrected coexpression score, ccs:

$$ccs = s/n_{res}$$

Iterative recalculation

The node in the network with the highest NDA score has, by definition, numerous strong correlations with other nodes in the subset R. The NDA scores assigned to these other nodes are therefore inflated by their association with the stongest node. This inflation may reflect biological reality, since both TSS have a putative genetic association with the phenotype of interest, and both share strong links. However, there is a risk that TSS sharing a chance association with a strongly coexpressed TSS will be spuriously inflated to significance. For this reason, we have applied a stringent correction in order to ensure that we have confidence in each significantly coexpressed TSS independently of all TSS with stronger coexpression in the network: the NDA score for each TSS is calculated after removing all TSS with stronger NDA scores from the network.

Input datasets

Of 267,225 robust promoters and enhancers identified by FANTOM5, 93,558 (50.6%) were promoters within 400 bases of the $5'$ end of a known transcript model. These were annotated with the name of the transcript. Alternative promoters were named in order of the highest transcriptional activity. Where necessary, coordinates for GWAS SNPs were translated to hg19 coordinates using LiftOver, or coordinates were obtained for SNP IDs from dbSNP version 138.

Permutations

A circular permutation method was devised to prevent systematic bias by maintaining the underlying structure of GWAS SNP data. The NDA score for a given regulatory region was compared with NDA scores obtained from randomly permuted subsets of genes to give an empirical p-value for coexpression. If permuted networks consist of randomly-selected regulatory regions, then this p-value quantifies coexpression alone; if the permuted networks are generated by mapping randomly-selected SNPs to regulatory regions, then the final p-value is a composite of two measures: coexpression, and the enrichment for true GWAS hits in regulatory sequence.

Pre-mapping permutations

Pre-mapping permutations use a random set of SNPs generated by rotation of the input set of SNPs, K, on a concatenated circular genome. The choice of background is critical - some more recent GWAS studies consider only a subset of variants with a high probability of association with a given trait. In the present analyses, background data were chosen to reflect as accurately as

possible the pool of variants included in the original study. For this reason, results are presented only for phenotypes for which the the entire summary dataset was available, including a p-value for every SNP, so that the background used to generate permuted networks is exactly the same background from which the real dataset is drawn.

Post-mapping permutations

In order to quantify the effect of coexpression alone (i.e. eliminating the inflation of NDA scores that occurs due to enrichment of trait-associated SNPs in regulatory regions), permuted networks were generated after mapping to TSS regions. This is analogous to randomly reassigning the labels in the network, but aims to preserve the local relationships between regulatory regions, since we cannot assume that regulatory regions are randomly distributed on the genome, and since regional regulatory events, such as chromatin reorganisation, are expected to lead to coexpression between nearby regulatory regions.

Where A is defined as a list of regulatory regions comprising the whole set of FANTOM5 TSS, post-mapping permutations select a subset of A in a similar circular manner, by displacing the members of the set R by a random number of places on the list. Where the displacement pushes members of R off the end of the list, they are re-entered at the beginning.

This process generates a pool of variants that are likely to be grouped in a similar distribution on the genome to the input set. If the input set contains a large group of TSS regions in close proximity to each other on the genome, it is likely that this group of TSS regions will be joined as a single unit (see above) for analysis. During generation of permutations, the same number of consecutive TSS regions elsewhere on the genome may not be in sufficient proximity (and expression correlation) to be grouped together. This would create extra network nodes, potentially inflating the NDA scores in the permuted sets. To mitigate against this, those TSS from each permutation that do not conform to the input set distribution are re-entered into a further circular permutation until an identical distribution is found. If no matching grouping is found after 8 repeat permutations, additional regulatory regions are added from consecutive positions above and below whichever group is nearest in size to the relevant group in the original input dataset.

False discovery rates (FDR) are calculated using the Benjamini-Hochberg method.

Choice of samples and regulatory regions

The enrichment for GWAS hits from a pooled resource comprising the NCBI GWAS catalog and the GWASdb database (observed $SNPs\ per\ Mb$: expected $SNPs\ per\ Mb$) was quantified at increasing search window sizes upstream and downstream from the transcription start site (TSS). A table of GWAS hits for a broad range of phenotypes was obtained from the NCBI GWAS catalog and from a larger, less selective catalog of GWAS p-values meeting permissive criteria for genome-wide significance, GWASdb. The GWASdb dataset is less curated than the NCBI GWAS catalog, but contains a much greater range of SNPs since it does not restrict inclusion to the strongest associations, or to putative causative variants. Because both databases are limited by the variation in reporting, and quality, of the original GWAS studies from which data are drawn, this analysis was restricted to variants meeting genome-wide significance at a widely-accepted threshold ($p < 5\times10^{-8}$). These catalogues were combined and filtered to remove duplicate entries. Data were obtained from:

- NHGRI GWAS catalog, June 2014 http://www.genome.gov/gwastudies

- GWASdb2, June 2014 update
  ftp://jjwanglab.org/GWASdb/20140629/gwasdb_20140629_snp_trait.gz

Overlapping phenotypes, such as “urate” and “uric acid” were manually merged. Phenotypes that were considered to be too broad to be informative were excluded, as were those that were not related to human disease. A complete table of phenotypes in GWASdb and NCBI GWAS catalog, showing mergers and inclusion/exclusion in the present work, is provided in a supplementary file (SF2_phenotype_matching.txt).

Anti-correlation

Strong anti-correlation between pairs of TSS associated with the same phenotype may have biological importance, such as down-regulation at one TSS but expression at another, or negative regulation of a signalling pathway on which expression of a TSS is dependent. For this reason, anti-correlations may improve detection of true associations in this analysis. However, in order to confer an overall improvement on the performance of the algorithm, true inverse expression relationships between phenotype-associated TSS would need to be sufficiently common to overcome the noise added by incorporating all strong anti-correlations into the NDA score. Anti-correlations do not contribute any net improvement to the NDA scores for a training set (Crohn’s disease, 50% of all SNPs, chosen at random), and were therefore excluded.

GWAS data sources

Full GWAS or meta-analysis data, reporting every SNP genotyped or imputed in a given study, are required in order to permute subsets against the appropriate background for a given study. These were obtained from the following sources:

- Crohn’s disease summary p-values were obtained from the International Inflammatory Bowel Disease Genetics Consortium ftp://ftp.sanger.ac.uk/pub4/ibdgenetics/cd-meta.txt.gz

- Ulcerative colitis summary p-values were obtained from the International Inflammatory Bowel Disease Genetics Consortium ftp://ftp.sanger.ac.uk/pub4/ibdgenetics/ucmeta-sumstats.txt.gz

- Summary p-values for human height were obtained from the GIANT consortium https://www.broadinstitute.org/collaboration/giant/images/4/47/GIANT_HEIGHT_LangoAllen2010_publicrelease_HapMapCeuFreq.txt

- Summary p-values for total cholesterol, LDL cholesterol, HDL cholesterol and triglycerides were obtained from the Global Lipids Consortium http://csg.sph.umich.edu/abecasis/public/lipids2013/

- Summary p-values for systolic and diastolic blood pressure. were obtained from the International Consortium on Blood Pressure study http://www.georgehretlab.org/icbp_088023401234-9812599.html

Cell type specificity

In order to better understand the pathophysiological implications of disease variants in regulatory regions, we sought to identify whether these regions exhibit unexpectedly specific expression in any given cell types or tissue samples. In order to reduce noise, technical and biological replicates were averaged for this and subsequent analyses. The full table of samples in FANTOM5, showing which samples were averaged as technical replicates, and which were excluded, is in supplementary table 2 (SF2_phenotype_matching.txt).

For a given trait, we took the subset of regulatory regions for which a significant coexpression pattern was detected for that trait (coexpression $FDR \leq 0.05$). For each regulatory region, we created a list of all cell types in which that region was active, ranked by expression level. We then combined the cell type lists for each regulatory region using a robust rank aggregation (RRA).

There are several possible sources of bias in this raw measurement. For example, some cell types have more cell-type specific transcriptional activity, perhaps because these cell types fulfil a specialised role; other cell types are particularly well-represented in the FANTOM5 samples. We therefore controlled for the probability that a given cell type would be highly ranked in the initial RRA analysis, by permuting RRA results for at least 100,000 random selections of n regulatory regions. We then calculated the empirical p-value for each cell type, i.e. the probability that this cell type would be assigined a raw RRA p-value at least as strong by random chance. We then corrected for multiple comparisons using the Benjamini-Hochberg method to estimate false discovery rate (FDR).

Code availability

Computer code required to run the NDA method, specifically for the detection of coexpression in FANTOM5 regulatory regions, can be obtained from https://github.com/baillielab/coexpression/

Results

Trait	SNPs included, p< 5×10^{-6} (SNPs per million bases)	Regulatory regions containing a SNP (SNPs per million bases)	Fold enrichment for SNPs in regulatory regions	Distinct regulatory regions	Significantly coexpressed TSS (FDR< 0.05)(% of distinct regions)	p(KS test)
Crohn's disease	1924 (0.6)	133 (3.5)	5.7	70	23 (33%)	1.61e-05
Ulcerative colitis	2162 (0.7)	146 (3.8)	5.5	83	20 (24%)	2.28e-06
LDL	4644 (1.5)	205 (5.2)	3.5	92	19 (21%)	1.48e-04
Total cholesterol	6421 (2.0)	316 (8.3)	4.1	128	29 (23%)	6.55e-07
Triglycerides	4863 (1.5)	254 (7.0)	4.6	97	23 (24%)	8.35e-06
Height	8882 (2.8)	358 (7.6)	2.7	166	29 (17%)	1.25e-06
HDL	5410 (1.7)	260 (7.2)	4.2	101	17 (17%)	3.51e-04
SBP	417 (0.1)	20 (0.4)	3.0	13	0 (0%)	4.89e-01
DBP	711 (0.2)	20 (0.4)	1.9	14	0 (0%)	5.41e-01

Table 1.

Evaluation of the NDA method and FANTOM5 input dataset

Our initial evaluation demonstrated that coexpression is stronger among regulatory regions containing variants with low GWAS p-values (Fig 2). The coexpression signal obtained for the test input set was evaluated using different subsets of FANTOM5 samples (cell lines, timecourses following a perturbation in primary cells or selected cell lines, tissue samples, primary cells, or various combinations of these), and different types of regulatory region (enhancers, promoters assigned to annotated genes, other promoters, or all regulatory regions combined) (Fig 3). The strongest coexpression is seen in the combined sample set. A "minimal detail" sample set was also tested, comprising a single average value for each of the timecourses, primary cell types and tissue types, and excluding data from unstimulated cell lines. The complete dataset, including all cell types and tissues, provided the strongest signal, demonstrating that there is additional biologically-relevant information contained in the expression profiles from all sample subsets (Fig 3).

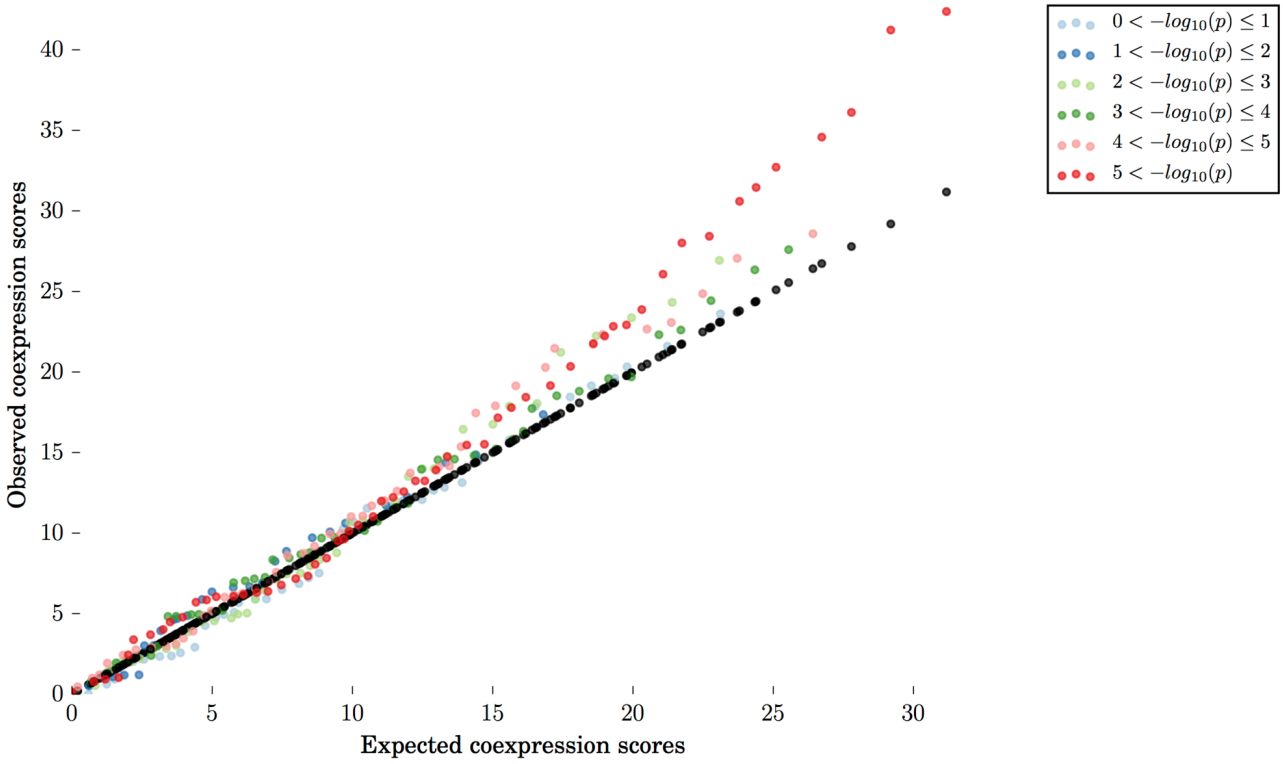

Fig 2. Optimisation of GWAS p-value threshold. Coexpression signals are shown for six different $-\log_{10}(p)$ bins for GWAS p values from a single study of Crohn's disease. From each bin, 800 SNPs were selected at random. No signal for coexpression is detected at weak p-values.

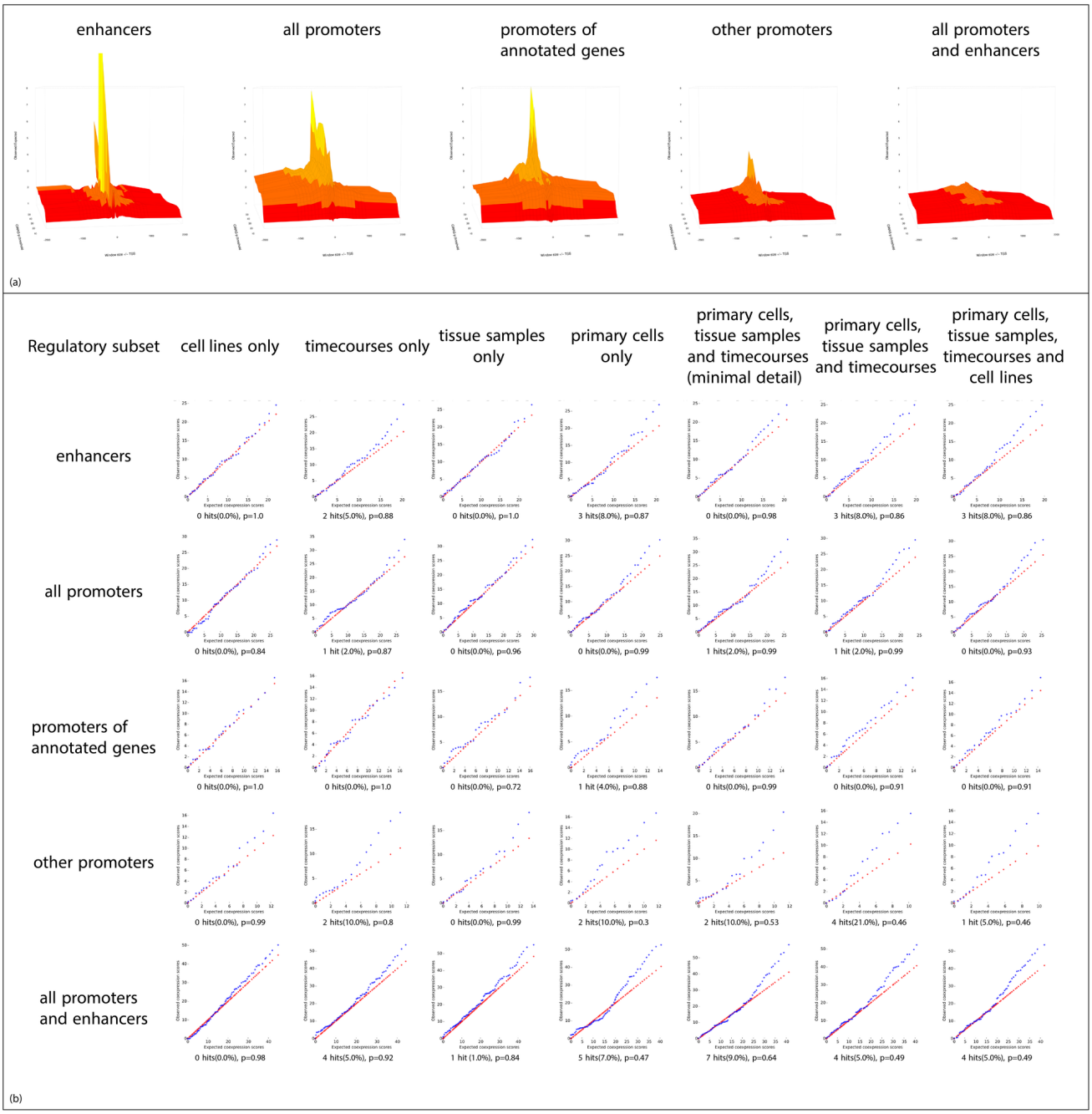

Fig 3. (a) Enrichment (y axis, observed $SNPs\ per\ Mb$: expected $SNPs\ per\ Mb$) at increasing search window sizes (x axis) upstream and downstream from the transcription start site (TSS) for increasingly strong GWAS signals (z axis, $-log_{10}p$). (b) Change in coexpression signal using different subsets of the FANTOM5 dataset, using the Crohn's disease GWAS as the input set. Q:Q plots of observed:expected NDA scores obtained using a given subset of samples (see methods for full description of each subset). Rows indicate the subset of regulatory regions used in each analysis. Percentage of significantly coexpressed entities (hits, $FDR < 0.05$) and p-value (Kolmogorov-Smirnov test) comparing observed (blue) and expected (red) distributions are shown below each plot.

The difference between the distributions of NDA scores derived from pre- and post-mapping permutations reveals the different components of the measure. When compared to a random pool of SNPs (pre-mapping permutations), two factors inflate the NDA scores for real GWAS data: firstly, more regulatory regions are identified because true GWAS hits are enriched within regulatory regions; secondly, the coexpression signal itself is greater for real data. In contrast, post-mapping permutations have precisely the same number of regulatory regions included as the real dataset, so there is no component of inflation due to enrichment in regulatory regions. The

effects of these different components are shown in Fig 4, which reveals the NDA score to be a composite measure of both signals.

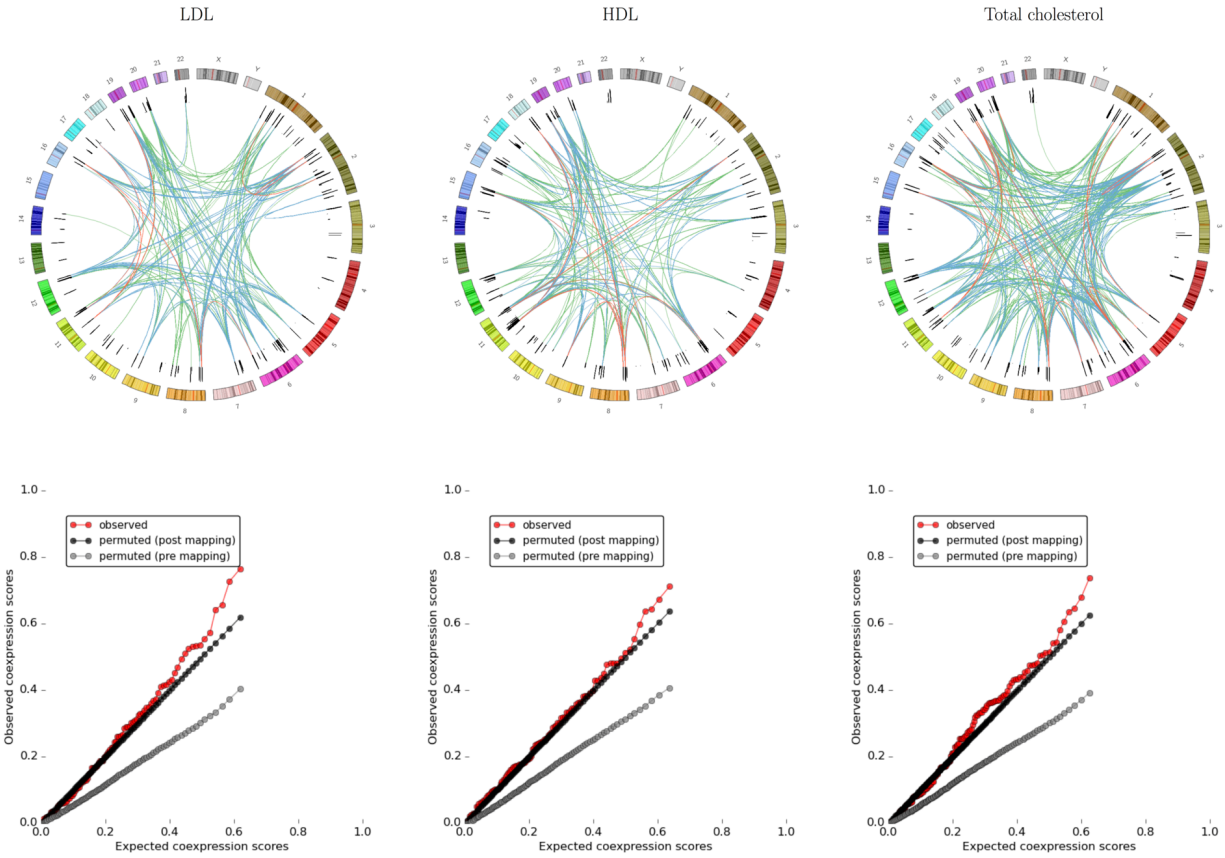

Fig 4. (Top panels) Circular plots of coexpression links between different locations on the genome, illustrating the spatial separation of highly-correlated regulatory regions. The coloured outer circle shows an end-to-end concatenated view of the human chromosomes. The black inner circle shows $-log_{10}$ GWAS p-values for included SNPs. Links depict an association between two regulatory regions containing these SNPs and are coloured according to $-log_{10}(p)$(line colour indicates $-log_{10}(p)$: red>3, blue> 2, green> 1.5). (Bottom panels) Quantile-quantile plots showing observed and expected coexpression scores. Expected coexpression scores are derived from circular permuted subsets of regulatory regions (post-mapping permutations; black circles) or SNPs chosen by circular permutations against the background of all SNPs genotyped in each study. Data are shown for high-density lipoprotein (HDL), low-density lipoprotein (LDL), and total cholesterol. See supplementary results for full results of all analyses.

Similar expression profiles are often seen arising from regulatory regions that are close to each other on the same chromosome, which may also span linkage disequilibrium blocks. The effect of this on the coexpression signal was mitigated by grouping nearby (within 100,000bp) regulatory regions into a single unit, unless they have notably different expression patterns. SNPs in nearby regulatory regions are also more likely to be in linkage disequilibrium, and these regulatory regions themselves are more likely to share cis- (or short-range trans-) regulatory signals in common. We checked for significant linkage disequilibrium between regulatory regions assigned to independent groups. At a threshold of $r^2 > 0.8$, there is no linkage disequilibrium between significantly coexpressed groups; three examples of weaker linkage relationships were detected with $0.08 \leq r^2 \leq 0.6$ (Supplementary results).

Fine mapping

Regulatory regions around individual TSS with higher coexpression scores contain variants with stronger GWAS p-values (Fig 5a), indicating that this independent signal provides additional information that may be used for fine-mapping causative loci (Fig 6; Supplementary results). Where data are available, we have compared our results to the recent fine mapping study by

Huang *et al*, who use high-resolution genotyping in 67,852 subjects with inflammatory bowel disease to quantify the probability that a given variant is causal. A total of 9 variants with a causal probability > 0.1 lie within 150,000bp of a significantly coexpressed region; of these, 7 lie immediately adjacent to the most significantly coexpressed promoter/enhancer in the region.

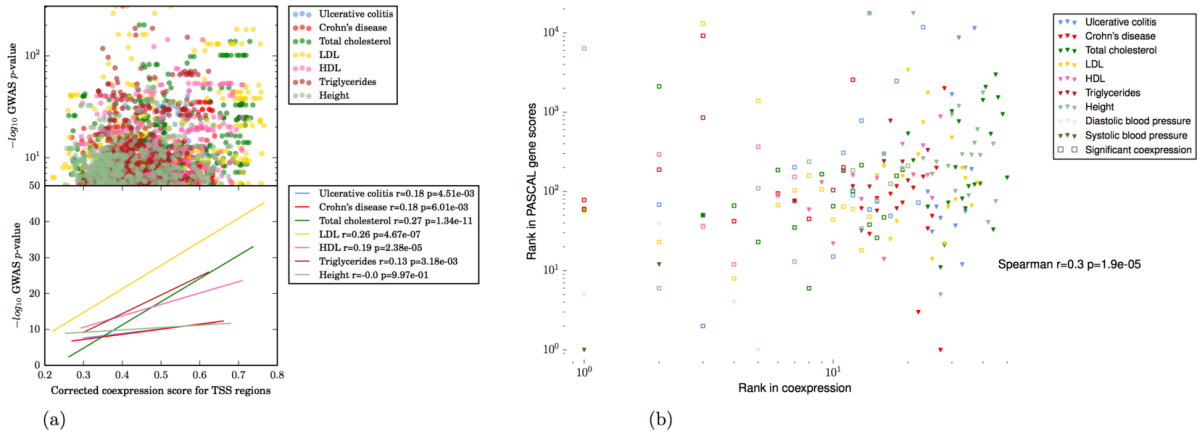

Fig 5. (a) Relationship between GWAS p-value for a SNP, and coexpression scores of individual promoters assigned to that SNP for all phenotypes for which significant coexpression was detected. Top panel: GWAS p-values (log scale) vs corrected coexpression scores. Bottom panel: linear regression lines for data in top panel; Spearman's r and associated p-values are shown for each trait. Only significantly coexpressed (FDR<0.05) promoters are included. (b) Rank comparison of named genes compared with gene-level burden of significance in original GWAS studies (PASCAL sum genescore). Log rank is shown on each axis (Rank 1 = highest scoring gene) for the subset of coexpression scores obtained for promoters of named genes. Open squares indicate significant coexpression (FDR<0.05).

Discovery and prioritisation of GWAS hits in regulatory sequence

In order to enable the detection of new regulatory regions with strong coexpression relationships, we chose a permissive threshold at GWAS $p < 5\times10^{-6}$. GWAS data for Crohn's disease[16] were used for initial optimisation of the NDA approach. Of the 8 GWAS datasets for phenotypes that were not used in algorithm development (i.e. all apart from Crohn's disease), 6 showed evidence of significant coexpression (Table 1). Among these, between 17 and 24% of regulatory regions identified as containing a GWAS SNP were found to be significantly coexpressed with other regulatory elements associated with the same phenotype (FDR < 0.05, compared with 100 permuted subsets of equal size; see Methods).

Table 1. Results of coexpression analysis for a range of human traits for which high-quality data are available: Crohn's disease, ulcerative colitis, high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol, triglycerides, height, systolic blood pressure (SBP) and diastolic blood pressure (DBP). KS test: Kolmogorov-Smirnov test comparing distribution of coexpression scores for this study with permuted values. *Initial optimisation and parameterisation of the algorithm was undertaken using a random subset of data from this study.

Although many coexpressed regulatory regions are not promoters for annotated genes (supplementary results; Fig 3), we compared the named genes in our results with gene-level burden of significance scores from PASCAL[17] analysis of the original GWAS studies. Since the coexpressed regulatory regions were detected due to the presence of a variant with a low p-value, it is expected that the genes with coexpressed promoters will be highly ranked in a gene-level analysis. However, the weak but significant correlation (Spearman $r = 0.30$; $p = 1.9\times10^{-5}$) between the approaches provides further evidence that the coexpression signal itself provides additional information which successfully prioritises regulatory regions (Fig 5b).

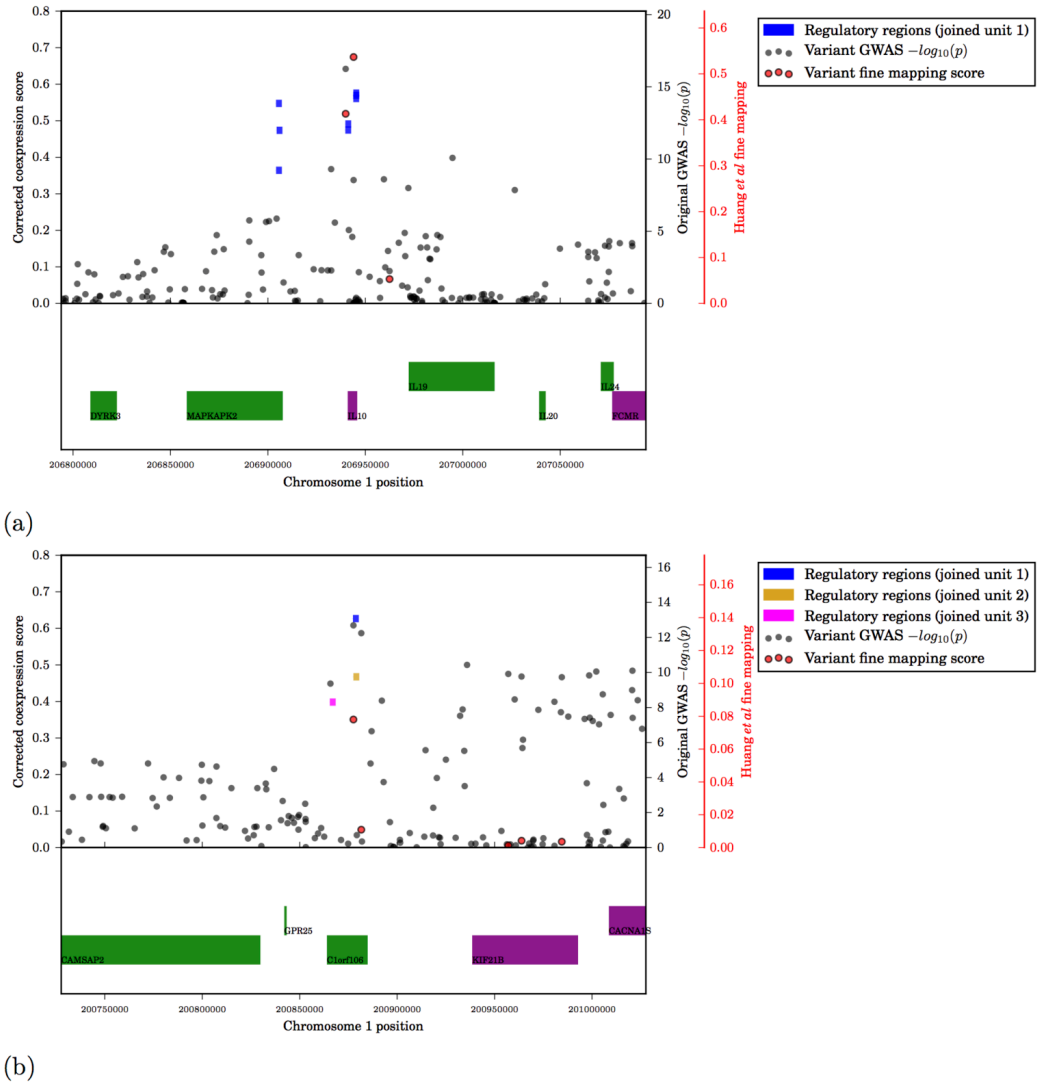

Fig 6. Examples of detail of chromosomal regions surrounding regulatory regions significantly coexpressed in ulcerative colitis (TSS+/-150Mb). (a) Region surrounding IL10 (b) Region surrounding C1orf106. Top panel: Coloured rectangles show genomic location of individual regulatory regions (promoters or enhancers). Height of regulatory regions on y-axis depicts the coexpression score assigned to this regulatory region; groups of regulatory regions considered as a single unit (see Methods) share the same colour. Black circles show GWAS p-values for individual SNPs. Red circles show causative probabilities estimated by Huang *et al* for specific variants, where available. Bottom panel: genomic locations of known protein coding transcripts in sense (green) and antisense (purple).

For a given disease, regulatory regions containing GWAS variants are coexpressed if they share similar activity patterns (i.e. similar expression patterns among transcripts arising from these regulatory regions) with other regulatory regions implicated in that disease. Fig 7a shows significant coexpression superimposed on a two-dimensional representation of the entire network of pairwise correlations. Since activity (transcript expression) was measured in many samples, the true proximity of regulatory regions to one another cannot be accurately represented in two dimensions – a perfect representation would require as many dimensions as there are unique samples. In contrast, the NDA method quantifies proximity of regulatory regions in true network space without artificial dimensionality reduction. Thus significantly coexpressed elements are

detected even if they are not directly adjacent on a two-dimensional representation of the network (Fig 7).

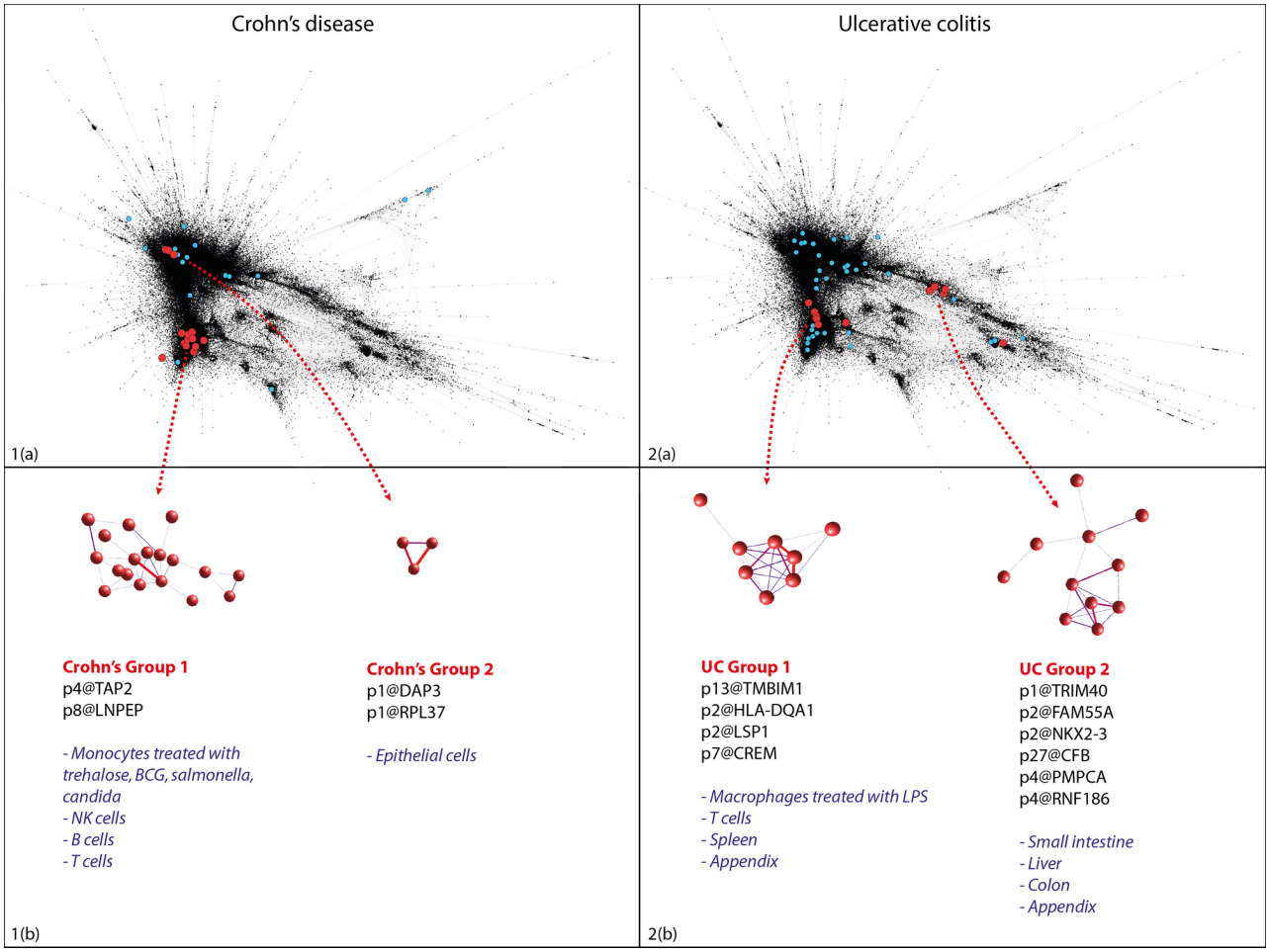

Fig 7. Examples of detail of chromosomal regions surrounding regulatory regions significantly coexpressed in ulcerative colitis ($TSS+/-150Mb$). (a) Region surrounding IL10 (b) Region surrounding C1orf106. Top panel: Coloured ectangles show genomic location of individual regulatory regions (promoters or enhancers). Height of regulatory regions on y-axis depicts the coexpression score assigned to this regulatory region; groups of regulatory regions considered as a single unit (see Methods) share the same colour. Black circles show GWAS p-values for individual SNPs. Red circles show causative probabilities estimated by Huang \emph{et al} for specific variants, where available. Bottom panel: genomic locations of known protein coding transcripts in sense (green) and antisense (purple).

We saw no evidence of spurious coexpression due to genomic proximity with shared regulatory influences (see Methods). In each of the GWAS analyses for which significant coexpression was detected, strong coexpression links were seen between loci that were widely separated on the genome (Fig 4; supplementary results).

The coexpression signal essentially combines the signal for association in a GWAS with the location and activity pattern of regulatory regions on the genome. We deliberately chose a permissive GWAS p-value threshold in order to enable the detection of new signals that did not achieve genome-wide significance in the original studies. For example, we found that coexpressed transcripts for both LDL and total cholesterol (TC) arise from promoters for well-studied genes such as APOB[18] and ABCG5[19], but also from regulatory regions not previously associated with cholesterol levels. A promoter for SLC22A1, which encodes an organic cation transporter, OCT1[20], is strongly coexpressed among elements associated with LDL and TC (Supplementary results). OCT1 transcription is regulated by cholesterol[21] and the transporter regulates hepatic steatosis through its role in thiamine transport[22]. This action of OCT1 is inhibited by metformin[22], an oral hypoglycaemic agent whose cholesterol-lowering effect[23] is not well

understood[24]. Full results of coexpression analyses are in the supplementary results, and online at http://baillielab.net/coexpression.

Cell-type and tissue specificity

The significantly-coexpressed networks detected here could be regarded as revealing the signature expression profile, at least within the FANTOM5 dataset, for a given disease or trait. We next explored whether these signature expression patterns reveal cell types or biological processes that may contribute to the trait or disease susceptibility.

We therefore ranked cell types and tissues by transcriptional activity for each of the significantly-coexpressed loci for each trait, and combined the rankings using a robust rank aggregation[25]. By first detecting the characteristic expression signature associated with a given phenotype using only high-resolution GWAS data, and then detecting the cell type and tissue activity profiles that underlie this signature, we improve on the statistical power of previous methods that have attempted to detect cell-type specific signatures of disease[4,6,26]. Signals that are strong enough to be detected in previous, less powerful studies are highly significant in our analysis; for example genetic loci associated with cholesterol are transcriptionally active in hepatocytes and liver tissue[6](Supplementary results).

Discussion

The development of high-throughput genotyping methods has led to an explosion of associations between genetic markers and human diseases[29]. The results presented here are a step towards overcoming the next challenge for this field: making sense of these associations to advance the practice of medicine. There has been increasing recognition of the potential to utilise prior knowledge to improve detection and interpretation of genome-wide signals[30]. The results of our analysis demonstrate that there is biological information in the coexpression of genetic variants associated with a particular disease that can provide the basis for prioritising variants that would not otherwise meet standard thresholds for genome-wide statistical significance.

We report relationships between numerous regulatory regions that are not associated with named genes – a restriction that has previously limited the transition from genetic discovery to biological understanding[14,31–34]. Our analysis reveals the impact of specific enhancers and promoters that may be remote from the genes they regulate, or may contribute to tissue-specific regulation of a gene that may otherwise appear to be more widely-expressed.

Even for those disease-associated variants that can be reliably assigned to a named gene, previous attempts to draw functional inferences have, by necessity, relied on published data[31], annotated biological pathways[35], or gene sets[34,36]. Although many important insights have been gained from these approaches, they share a fundamental limitation: reliance on existing knowledge. This restricts the ability to exploit the potential of genomics to deliver insights into new, previously unseen, mechanisms of disease[37].

Results for Crohn's disease and ulcerative colitis were compared to the report by Huang *et al[15]*, who used high-resolution genotyping in a large cohort, together with publicly-available functional genomics data, to identify immune cell signatures implicated in Crohn's disease, and gut-specific cell types in ulcerative colitis. Our analysis, conducted in parallel and without knowledge of these findings, discovered the same associations, but goes further. Firstly, we demonstrate with a higher level of statistical confidence that these cell type associations are real (supplementary results). This is important in itself, because it is consistent with the view that ulcerative colitis, in which disease processes are primarily restricted to the colon and rectum, is a consequence of

dysregulation of processes that are intrinsic to the large bowel, including epithelial barrier function[27], whereas Crohn's disease is a multisystem autoimmune disorder with more diverse extra-intestinal manifestations[28], consistent with a primary innate immune aetiology affecting monocyte-macrophage differentiation and response to micro-organisms[38].

Secondly, our analysis extends current knowledge by revealing two distinct groups of significantly-coexpressed regulatory regions for each of these diseases, with differing expression profiles. For Crohn's disease, one group is restricted to immune cells, particularly monocytes exposed to inflammatory stimuli, while another group of regulatory regions is active in epithelial cells. In contrast, cell type associations with ulcerative colitis were statistically significant in rectum, colon and intestine samples, and in a distinct group of immune cells: macrophages exposed to bacterial lipopolysaccharide (Fig 7; S5 Table 1.2). Based on the fundamental assumption of coexpression - that expression profile relates to function - we interpret this as evidence that two distinct biological processes are implicated in each of these diseases. This may be because a "two-hit" mechanism is required for disease pathogenesis. Alternatively these distinct processes may indicate genetically-(and hence aetiologically-) distinct sub-syndromes, or disease endotypes[39], within both Crohn's disease and ulcerative colitis.

In either case the predominance of each process in an individual patient is likely to have therapeutic relevance. For example, the highly variable clinical response to immunomodulatory therapies, such as methotrexate[40] or anti-TNF monoclonal antibodies[41], may be influenced by the burden of disease-associated variants in Crohn's disease Group 1 (Fig 7). This represents a conceptually new application of network theory to the detection of disease endotypes, and is likely to have more direct clinical consequences than other methods[42].

The data used for development and testing of the coexpression approach were from large meta-analyses that incorporate genotyping (or imputation) of genetic variants at extremely high resolution, increasing the probability that variants will be found within regulatory regions. In future, the availability of whole-genome sequencing can reasonably be expected to produce many additional high-quality datasets for coexpression analysis. In principle, the NDA approach can be generalised to any network in which it is desirable to quantify the proximity of a subset of nodes.

The scale, depth and breadth of the FANTOM5 expression atlas enable detection of subtle coexpression signals for regulatory regions that have previously been undetectable. The NDA approach developed here enables the identification of cell types and regulatory regions implicated in disease pathogenesis, and contributes a new independent signal to fine mapping of causative loci. As additional genetic studies become available at greater genotyping resolution, we anticipate that this method will detect new genetic associations with disease and coexpressed modules underlying pathogenesis. The NDA method will enable the identification of critical cell types and processes implicated in mechanisms of disease, and enable further genetic stratification of disease endotypes by underlying mechanism.

DATA ACCESS

The FANTOM5 atlas is accessible from http://fantom.gsc.riken.jp/data/

An online service running the coexpression method is available at http://baillielab.net/coexpression

SUPPORTING INFORMATION LEGENDS

1. SF1_specific_GWAS_results.xlsx Network density analysis results for each of the GWAS studies included here.
2. SF2_phenotype_matching.txt Table of phenotypes in GWAS catalog and GWASdb showing combinations of similar/identical phenotypes for use in regional enrichment calculations.
3. SF3_sample_averaging.xlsx Table of FANTOM5 samples showing combinations of similar samples that were averaged for use in coexpression network.
4. SF4_supplementary_fine_mapping_results.pdf Figures for every region included in the analyses described in this paper (See Fig 6 in main manuscript for legend).
5. SF5_supplementary_results.pdf Tables of network density analysis results for each GWAS studies included here.

REFERENCES

1. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nat Rev Genet. 2013;14: 139–149. doi:10.1038/nrg3377
2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337: 1190–1195. doi:10.1126/science.1222794
3. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507: 455–461. doi:10.1038/nature12787
4. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518: 337–343. doi:10.1038/nature13835
5. Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J.K., et al. A promoter-level mammalian expression atlas. Nature. 2014;507: 462–470. doi:10.1038/nature13182
6. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat Meth. 2016;13: 366–370. doi:10.1038/nmeth.3799
7. The FANTOM Consortium, Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. The Transcriptional Landscape of the Mammalian Genome. Science. 2005;309: 1559–1563. doi:10.1126/science.1112014

8. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015;347: 1010–1014. doi:10.1126/science.1259418

9. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47: 1228–1235. doi:10.1038/ng.3404

10. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional category using GWAS summary statistics. bioRxiv. 2015; 014241. doi:10.1101/014241

11. Hume DA, Summers KM, Raza S, Baillie JK, Freeman TC. Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations. Genomics. 2010;95: 328–338. doi:10.1016/j.ygeno.2010.03.002

12. Mabbott NA, Baillie JK, Hume DA, Freeman TC. Meta-analysis of lineage-specific gene expression signatures in mouse leukocyte populations. Immunobiology. 2010;215: 724–736. doi:10.1016/j.imbio.2010.05.012

13. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47: 569–576. doi:10.1038/ng.3259

14. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015;6: 5890. doi:10.1038/ncomms6890

15. Huang H, Fang M, Jostins L, Umićević Mirkov M, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;547: 173–178. doi:10.1038/nature22969

16. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42: 1118–1125. doi:10.1038/ng.717

17. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLOS Comput Biol. 2016;12: e1004714. doi:10.1371/journal.pcbi.1004714

18. Tybjærg-Hansen A, Steffensen R, Meinertz H, Schnohr P, Nordestgaard BG. Association of Mutations in the Apolipoprotein B Gene with Hypercholesterolemia and the Risk of Ischemic Heart Disease. N Engl J Med. 1998;338: 1577–1584. doi:10.1056/NEJM199805283382203

19. Lee M-H, Lu K, Hazard S, Yu H, Shulenin S, Hidaka H, et al. Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. Nat Genet. 2001;27: 79–83. doi:10.1038/83799

20. Klaassen CD, Aleksunes LM. Xenobiotic, Bile Acid, and Cholesterol Transporters: Function and Regulation. Pharmacol Rev. 2010;62: 1–96. doi:10.1124/pr.109.002014

21. Dias V, Ribeiro V. The expression of the solute carriers NTCP and OCT-1 is regulated by cholesterol in HepG2 cells. Fundam Clin Pharmacol. 2007;21: 445–450. doi:10.1111/j.1472-8206.2007.00517.x

22. Chen L, Shu Y, Liang X, Chen EC, Yee SW, Zur AA, et al. OCT1 is a high-capacity thiamine transporter that regulates hepatic steatosis and is a target of metformin. Proc Natl Acad Sci U S A. 2014;111: 9983–9988. doi:10.1073/pnas.1314939111

23. Bailey CJ, Turner RC. Metformin. N Engl J Med. 1996;334: 574–579. doi:10.1056/NEJM199602293340906

24. Shaw RJ, Lamia KA, Vasquez D, Koo S-H, Bardeesy N, Depinho RA, et al. The kinase LKB1 mediates glucose homeostasis in liver and therapeutic effects of metformin. Science. 2005;310: 1642–1646. doi:10.1126/science.1120781

25. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012;28: 573–580. doi:10.1093/bioinformatics/btr709

26. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473: 43–49. doi:10.1038/nature09906

27. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007;448: 427–434. doi:10.1038/nature06005

28. Mekhjian HS, Switz DM, Melnyk CS, Rankin GB, Brooks RK. Clinical features and natural history of Crohn's disease. Gastroenterology. 1979;77: 898–906.

29. Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 2012;40: D1047–D1054. doi:10.1093/nar/gkr1182

30. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17: 144. doi:10.1186/s12864-016-2443-6

31. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, International Schizophrenia Consortium, Purcell SM, et al. Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. Storey JD, editor. PLoS Genet. 2009;5: e1000534. doi:10.1371/journal.pgen.1000534

32. Wojcik GL, Kao WL, Duggal P. Relative performance of gene- and pathway-level methods as secondary analyses for genome-wide association studies. BMC Genet. 2015;16. doi:10.1186/s12863-015-0191-2

33. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet. 2011;7: e1001273. doi:10.1371/journal.pgen.1001273

34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102

35. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

36. Nam D, Kim J, Kim S-Y, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. Nucl Acids Res. 2010;38: W749–W754. doi:10.1093/nar/gkq428

37. Baillie JK. Targeting the host immune response to fight infection. Science. 2014;344: 807–808. doi:10.1126/science.1255074

38. Baillie JK, Arner E, Daub C, Hoon MD, Itoh M, Kawaji H, et al. Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease. PLOS Genetics. 2017;13: e1006641. doi:10.1371/journal.pgen.1006641

39. Russell CD, Baillie JK. Treatable traits and therapeutic targets: Goals for systems biology in infectious disease. Current Opinion in Systems Biology. 2017;2: 139–145. doi:10.1016/j.coisb.2017.04.003

40. McDonald JWD, Wang Y, Tsoulis DJ, MacDonald JK, Feagan BG. Methotrexate for induction of remission in refractory Crohn's disease. Cochrane Database Syst Rev. 2014; CD003459. doi:10.1002/14651858.CD003459.pub4

41. Peyrin-Biroulet L, Deltenre P, de Suray N, Branche J, Sandborn WJ, Colombel J-F. Efficacy and safety of tumor necrosis factor antagonists in Crohn's disease: meta-analysis of placebo-controlled trials. Clin Gastroenterol Hepatol. 2008;6: 644–653. doi:10.1016/j.cgh.2008.03.014

42. Hall DP, MacCormick IJC, Phythian-Adams AT, Rzechorzek NM, Hope-Jones D, Cosens S, et al. Network Analysis Reveals Distinct Clinical Syndromes Underlying Acute Mountain Sickness. PLoS ONE. 2014;9: e81229. doi:10.1371/journal.pone.0081229